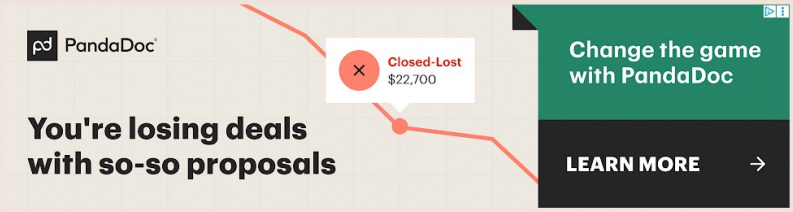
EXHIBIT 178



BUSINESS + TECH +

Exclusive: Anthropic Drops Flagship Safety Pledge

ADD TIME ON GOOGLE



by Billy Perrigo
CORRESPONDENT

FEB 24, 2026 3:00 PM ET



This photograph shows a smartphone screen displaying the logo of Anthropic, an American company specializing in artificial intelligence (AI). It develops Claude, a family of large language models, and is also known for its research in the field of AI safety in Creteil, France, on September 9, 2025. *Samuel Boivin/NurPhoto via Getty Images*

Anthropic, the wildly successful AI company that has cast itself as the most safety-conscious of the top research labs, is dropping the central pledge of its flagship safety policy, company officials tell TIME.

In 2023, Anthropic committed to never train an AI system unless it could guarantee in advance that the company's safety measures were adequate. For years, its leaders touted that promise—the central pillar of their Responsible Scaling Policy (RSP)—as evidence that they are a responsible company that would withstand market incentives to rush to develop a potentially dangerous technology.

Recommended Stories


The U.S. Supreme Court Will Hear a Major Climate Change Case This Fall


Taiwan's Opposition Leader Says Talking to China Is the Island's Best Defense


More Than One Third of U.S. Counties Are Maternity Care Deserts

But in recent months the company decided to radically overhaul the RSP. That decision included scrapping the promise to not release AI models if Anthropic can't guarantee proper risk mitigations in advance.

"We felt that it wouldn't actually help anyone for us to stop training AI models," Anthropic's chief science officer Jared Kaplan told TIME in an exclusive interview. "We didn't really feel, with the rapid advance of AI, that it made sense for us to make unilateral commitments ... if competitors are blazing ahead."

The new version of the policy, which TIME reviewed, includes commitments to be more transparent about the safety risks of AI, including making additional disclosures about how Anthropic's own models fare in safety testing. It commits to matching or surpassing the safety efforts of competitors. And it promises to "delay" Anthropic's AI development if leaders both consider Anthropic to be leader of the AI race and think the risks of catastrophe to be significant.





But overall, the change to the RSP leaves Anthropic far less constrained by its own safety policies, which previously categorically barred it from training models above a certain level if appropriate safety measures weren't already in place.

The change comes as Anthropic, previously considered to be behind OpenAI in the AI race, rides the high of a string of technological and commercial successes. Its Claude models, especially the software-writing tool Claude Code, have won legions of devoted fans. In February, Anthropic raised $30 billion in new investments, valuing it at some $380 billion, and reported that its annualized revenue was growing at a rate of 10x per year. The company's core business model of selling direct to businesses is seen by many investors as more credible than OpenAI's main strategy of monetizing a vast consumer user base.

Advertisement



Kaplan, the Anthropic executive and co-founder, denied the company's decision to change course was a capitulation to market incentives as the race for superintelligence accelerates. He framed it instead as a pragmatic response to emerging political and scientific realities. "I don't think we're making any kind of U-turn," Kaplan says.

When Anthropic introduced the RSP in 2023, Kaplan says, the company hoped it would encourage rivals to adopt similar measures. (No rivals made quite as overt a promise to pause AI development, but many published lengthy reports detailing their plans to mitigate risk, which Kaplan chalks up as Anthropic exerting a good influence on the industry.) Executives also hoped the approach might eventually serve as a blueprint for binding national regulations or even international treaties, Kaplan claims.

But those regulations never materialized. Instead, the Trump Administration has endorsed a let-it-rip attitude to AI development, even going so far as to attempt to nullify state regulations. No federal AI law is on the horizon. And while a global governance framework may have seemed possible in 2023, three years later it has become clear that door has closed. Meanwhile, competition for AI supremacy—between companies but also between nations—has only intensified.

Advertisement



To make matters worse, the science of AI evaluations has proven more complicated than Anthropic expected when it first crafted the RSP. The arrival of powerful new models meant that, in 2025, Anthropic announced it could not rule out the

possibility of these models facilitating a bio-terrorist attack. But while they couldn't rule it out, they also lacked strong scientific evidence that models *did* pose that kind of danger, which made it difficult to convince governments and rivals of what they saw as the need to act carefully. What the company had previously imagined might look like a bright red line was instead coming into focus as a fuzzy gradient.

For nearly a year, Anthropic executives discussed ways to reshape their flagship safety policy to match this new environment, Kaplan says. One point they kept coming back to was their founding premise: the idea that to do proper AI safety research, they had to build models at the frontier of capability—even though doing so might accelerate the arrival of the dangers they feared.

Advertisement



In February, according to Kaplan, Amodei decided that keeping the company from training new models while competitors raced ahead would be helpful to nobody. "If one AI developer paused development to implement safety measures while others moved forward training and deploying AI systems without strong mitigations, that could result in a world that is less safe," the new version of the RSP, approved unanimously by Amodei and Anthropic's board, states in its introduction. "The developers with the weakest protections would set the pace, and responsible developers would lose their ability to do safety research."

Chris Painter, the director of policy at METR, a nonprofit focused on evaluating AI models for risky behavior, reviewed an early draft of the policy with Anthropic's permission. He says the change is understandable — but also a bearish signal for the world's ability to navigate potential AI catastrophes. The change to the RSP shows Anthropic "believes it needs to shift into triage mode with its safety plans, because methods to assess and mitigate risk are not keeping up with the pace of capabilities," Painter tells TIME. "This is more evidence that society is not prepared for the potential catastrophic risks posed by AI."

Advertisement



Anthropic argues the retooled RSP is designed to keep the biggest benefits of the old one. For example, by constraining itself from releasing new models, Anthropic's original RSP also incentivized it to quickly build safety mitigations. (Because otherwise the company would be unable to sell its AI to customers.) Anthropic says it believes it can maintain that incentive. The new policy commits the company to regularly release what it calls "Frontier Safety Roadmaps": documents laying out a list of detailed goals for future safety measures it hopes to build.

"We hope to create a forcing function for work that would otherwise be challenging to appropriately prioritize and resource, as it requires collaboration (and in some cases sacrifices) from multiple parts of the company and can be at cross-purposes with immediate competitive and commercial priorities," the new RSP states.

Anthropic says it will also commit to publishing so-called "Risk Reports" every three to six months. The reports, the company says, will "explain how capabilities, threat models (the specific ways that models might pose threats), and active risk mitigations fit together, and provide an assessment of the overall level of risk." These documents will be more in-depth than the reports the company already publishes, a spokesperson tells TIME.

BY KARGO



Advertisement



"I like the emphasis on transparent risk reporting and publicly verifiable safety roadmaps," says Painter, the METR policy official. But he said he was "concerned" that moving away from binary thresholds under the previous RSP, by which the arrival of a certain capability could act as a tripwire to temporarily halt Anthropic's AI development, might enable a "frog-boiling" effect, where danger slowly ramps up without a single moment that sets off alarms.

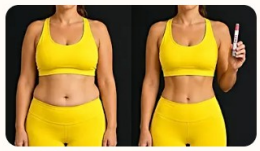
Asked whether Anthropic was caving to market pressure, Kaplan argued that, in fact, Anthropic was making a renewed commitment to developing AI safely. "If all of our competitors are transparently doing the right thing when it comes to catastrophic risk, we are committed to doing as well or better," he said. "But we don't think it makes sense for us to stop engaging with AI research, AI safety, and most likely lose relevance as an innovator who understands the frontier of the technology, in a scenario where others are going ahead and we're not actually contributing any additional risk to the ecosystem."

Advertisement





Forbes Reveals Clear GLP-1 Winner 2026 (See prices)
Lowest price ever - Lock in $79/month
Forbes | Sponsored
Learn More



Forbes Reveals Best GLP-1 Medications of 2026
Lowest price ever - Lock in $79/month
Forbes | Sponsored
Learn More

TIME

Books: Major Minor Poet
John Betjeman, 52, is a gentle, witty, rumpled Englishman who has been called "the greatest bad poet now living." It would be in character if he agreed with that estimate, ...
TIME



Spinal Stenosis Is Not From Tight Muscles. Meet The Real Enemy of...
Smoothspine | Sponsored
Learn More



Temu Clearance Sale
Temu Clearance Sale, where unbeatable discounts meet your shopping needs
TEMU | Sponsored
Shop Now



Legendary Stock Bull Says: “Brace for an Epic Price Crash”

Alexander Green believes we’re on the verge of perhaps the greatest “Price Crash” in more than a century. And if history repeats, it could be one of the best things that ever ...

The Oxford Club | Sponsored

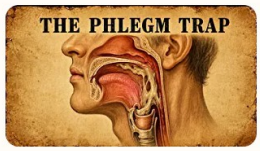
Learn More



New York: 24 Hour Dental Implants

Take our 1-minute quiz to find out if you're eligible for our revolutionary 24-hour ...

nuviasmiles | Sponsored



That "mucus" in your throat may not be mucus at all

The constant throat mucus and need to clear it often isn't phlegm at all it can be ...

NanoRevive | Sponsored

Learn More



$149 Tirzepatide: One Flat Price, Every Dose, Every Month

Doing everything right but still stuck? Here's the metabolic reason plateaus happen - and how compounded tirzepatide can help, starting at $149/mo.

CoreAge Rx | Sponsored

Learn More

TIME

Game Changers - TIME

As co-founders of Warby Parker, Neil Blumenthal and Dave Gilboa shake up the eyewear industry to provide affordable, stylish, high-quality glasses for people all over the world

TIME



Costco Shoppers Say This Patch Helped Them Fit into Their Old Jeans (It's Genius!)

Doctors Say It Targets the Hormone Behind Constant Hunger and Cravings and Costumers are Already Noticing the Impact.

health-trends-Journal.com | Sponsored



12 Expenses Retirees Are Quietly Dropping

Greensprout | Sponsored



The Worst Zip Codes for Car Insurance in New York (Is Yours on...

Check rates now—many drivers are paying more than they need to.

OTTO Insurance | Sponsored

Click Here



Meet Will Scharf, Trump's New White House Counsel Who Shares an Interesting History With the President

As Scharf prepares to take on new responsibilities, here's what to know about his history with Trump, his political aspirations, and more.

TIME



Japan's Oldest Doctors Say Cognitive Decline Isn't Age: Just Stop Eating These 3 Foods

slowvitalco | Sponsored

Learn More



Top 10 Ridiculously Strong Drinks - TIME

TIME



Why Men Wear This in Private — and Women Are Left Gasping

My Husband Started Wearing This Black Bracelet. I Stopped Asking Questions and Started Smiling

The MojoFlow | Sponsored

Learn More





Summer fun just got safer with the ADT Summer Sale

Focus on your summer fun. With our 24/7 pro monitoring, we've got your back when every second counts. Get your system during the ADT Summer Sale.

ADT | Sponsored

Learn More